Cell PRESS
Structure
Article

Probing the "Dark Matter" of Protein Fold Space

William R. Taylor,[1,3,*] Vijayalakshmi Chelliah,[1] Siv Midtun Hollup,[3] James T. MacDonald,[1] and Inge Jonassen[2,3]
[1]Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK
[2]Computational Biology Unit
[3]Department of Informatics
University of Bergen, Norway
*Correspondence: wtaylor@nimr.mrc.ac.uk
DOI 10.1016/j.str.2009.07.012

SUMMARY

We used a protein structure prediction method to generate a variety of folds as α-carbon models with realistic secondary structures and good hydrophobic packing. The prediction method used only idealized constructs that are not based on known protein structures or fragments of them, producing an unbiased distribution. Model and native fold comparison used a topology-based method as superposition can only be relied on in similar structures. When all the models were compared to a nonredundant set of all known structures, only one-in-ten were found to have a match. This large excess of novel folds was associated with each protein probe and if true in general, implies that the space of possible folds is larger than the space of realized folds, in much the same way that sequence-space is larger than fold-space. The large excess of novel folds exhibited no unusual properties and has been likened to cosmological dark matter.

INTRODUCTION

With growing numbers of known protein structures being determined, the occurrence of a novel fold for a globular protein of reasonable size (over 100 residues) is becoming an increasingly rare event (Grant et al., 2004). This has led to speculation that we are close to having a structural representation for every basic protein fold (Zhang and Skolnick, 2005a; Taylor, 2007b), with the implication that all natural protein sequences can be constructed from known structures through the assembly of domains taken from the current collection. This position may have been reached either because there is, in principle, only a limited number of possible protein structures or by historical accident in which nature has been "lazy" and restricted herself to the reuse of a limited set of protein folds. Starting from a few basic folds, the known evolutionary mechanisms of gene duplication, fusion, and deletion could have generated the current variety of protein structures. However, this is difficult to prove because specific sequence and structural similarities become lost over such a long time span. It is easier to investigate the alternative possibility that there is a theoretical constraint on the number of possible folds. If this exists and imposes a limit near to the number of known folds, then it can be inferred that the boundaries of observed fold space are probably constrained. If, however, there is a large or no apparent limit, then the folds we see are likely to be constrained mainly by their history.

In principle, such a computational experiment could be carried out by generating every possible protein sequence (or a large number of them) and predicting the three-dimensional (3D) structure for each sequence. Even without this being an enormous calculation, methods for the ab initio prediction of structure are not sufficiently accurate for any moderately sized protein. Small structures (up to 100 residues) can be predicted with reasonable accuracy using fragments of known structures, but given the nature of the current question, the use of known structures would result in a biased approach and, in addition, the variety of fold complexity for small proteins is limited. We have therefore taken a different approach to test this hypothesis, and to avoid the use of known structures. We have adapted a structure prediction method based on an idealized secondary structure lattice representation that constructs realistic models of protein structures (Taylor et al., 2008).

Given the specification of just the order of secondary structure type (α or β), this approach allows a wide variety of protein folds to be explored through the combinatorial enumeration of all paths connecting different points in a secondary structure lattice (Taylor, 2002). In the generation of these folds, generic principles of protein structure, such as handedness of connection and loop-crossing, provide constraints that reduce the number of folds constructed from millions to several thousand. Because each of these models is derived from a defined lattice, it is possible to encode their fold as a simple string, referred to as a "topology string".[1] This representation provides a fast way to select unique folds and compare them with other models and native structures represented in the same way. It is this automatic topological definition of a protein fold that allows us to rigorously define when two structures have a topological match that would be undetected or wrongly identified by root-mean-square deviation (rmsd)-based superposition methods of

[1]We use the terms "fold" and "topology" interchangeably. The term "topology" is used more loosely than its strict mathematical counterpart but embodies the equivalent concept that proteins with the same topology may be twisted or sheared but still retain the same relative packing and orientation of secondary structure elements. There is therefore a unique one-to-one mapping between a fold and its topology string. Our concept of a fold is purely geometrical and has no connection with function.

comparison. Some might view such an approach as overly restrictive, but without a precise definition of a protein fold at the topological level, any analysis of these large data sets would be difficult, if not impossible. (See Experimental Procedures for string encoding and matching details).

Given the simplicity of our representation, it has been possible to generate models for proteins that are large enough to exhibit nontrivial topologies and even knots. Using this approach, we show that many novel folds can be constructed, suggesting that nature has made use of only a very limited subset of the possible protein folds.

RESULTS

To generate a large population of protein folds, we took five medium-sized proteins (all over 100 residues) of the three-layer beta-alpha class and generated variation around their known secondary structure. This was done by selecting subsets of each sequence family, making a multiple sequence alignment and predicting the secondary structures. Because these prediction methods attain 80% accuracy at best, considerable variation was introduced into the number and location of the secondary structure elements, both within and between the different alignments. Additional variation was also introduced by permuting any ambiguous secondary structure elements to all combinations of alpha and beta structure state. The predictions, based on ten different multiple sequence alignments for five different proteins, were then processed through the modeling protocol and the resulting models ranked on a score that reflected how well the hydrophobic core was packed and the degree to which the predicted and modeled secondary structures were colocated (referred to as the "model quality score"). By this measure, the highest-scoring fold with a unique topology string was scanned against the known structures and the number of matches recorded.

An important confirmation that realistic models have been generated is whether the population of models contains the native fold of the proteins that were used in the initial multiple sequence alignments. In each run, the top ten ranked models typically included the native fold, showing that the model quality score selected structures with native protein-like features. Among the other high-scoring folds were several native-like folds with rearrangements of β strands to nearby positions in the sheet (strand-swapping) and of these, on average, half had found matches to known folds. Across the remainder, of typically 100 folds there was a sparse but relatively even scattering of matches to known folds with only some reduction in matches toward the tail of the lists.

Typically, the range of folds generated from each alignment had only 50% overlap, so when these results were pooled, the number of folds was greatly expanded, producing 2148 distinct topology strings of which only 132 were found in known structures. In other words, starting from just five probe proteins, only 6% of the protein-like folds we generated correspond to those found in nature. This means that, on average, we found 400 novel folds associated with each probe fold in our test set. By analogy with unseen components in the cosmological distribution of matter, we refer to these unobserved folds as the "dark matter" of protein fold space. In Figure 1, an example is shown of a model with a well-known topology, corresponding to 29 folds in the nonredundant Protein Data Bank (PDB) and next to it is an example of a novel fold. In Figure 2, both novel and known folds obtained from one probe protein are plotted in an arbitrary 3D space using multidimensional scaling to show their relationships.

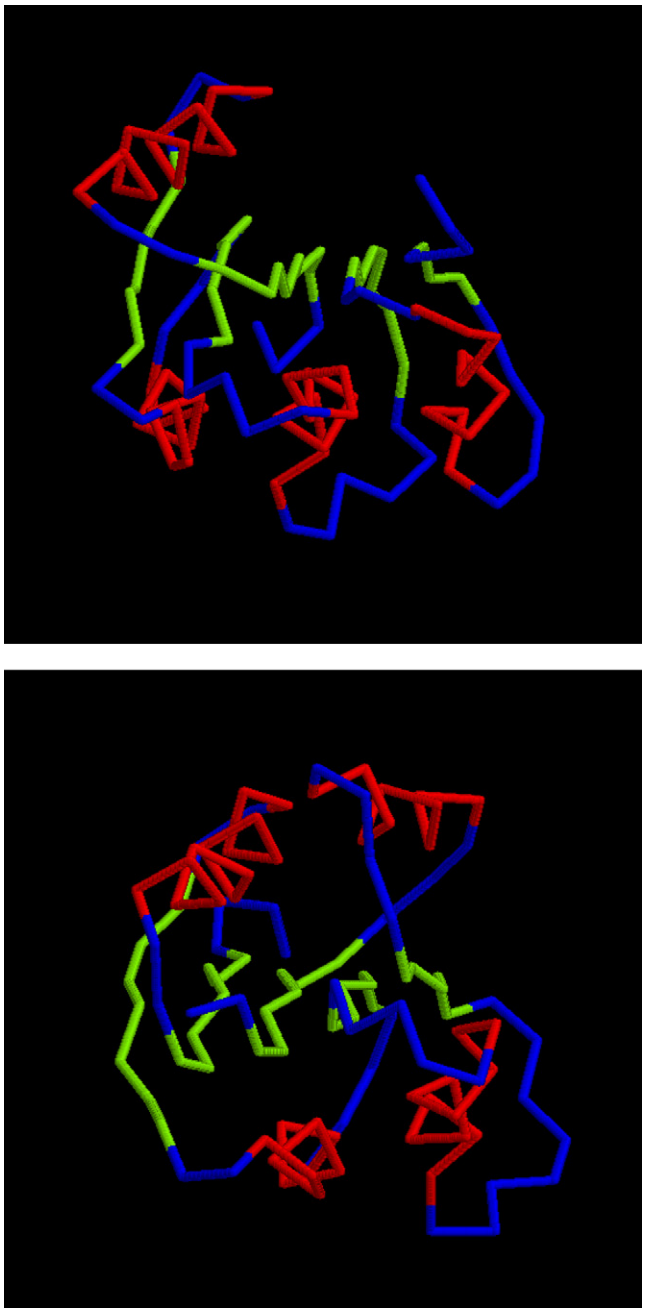


Figure 1. Examples of Models with a Known and Novel Fold
Two α-carbon-based models are shown with α helices colored red and β strands green. Fold *a* is found in 29 known structures whereas fold *b* corresponds to none. The representation of these folds as topology strings is illustrated in Experimental Procedures (Figure 4) along with their comparison with known structures (Figure 5).

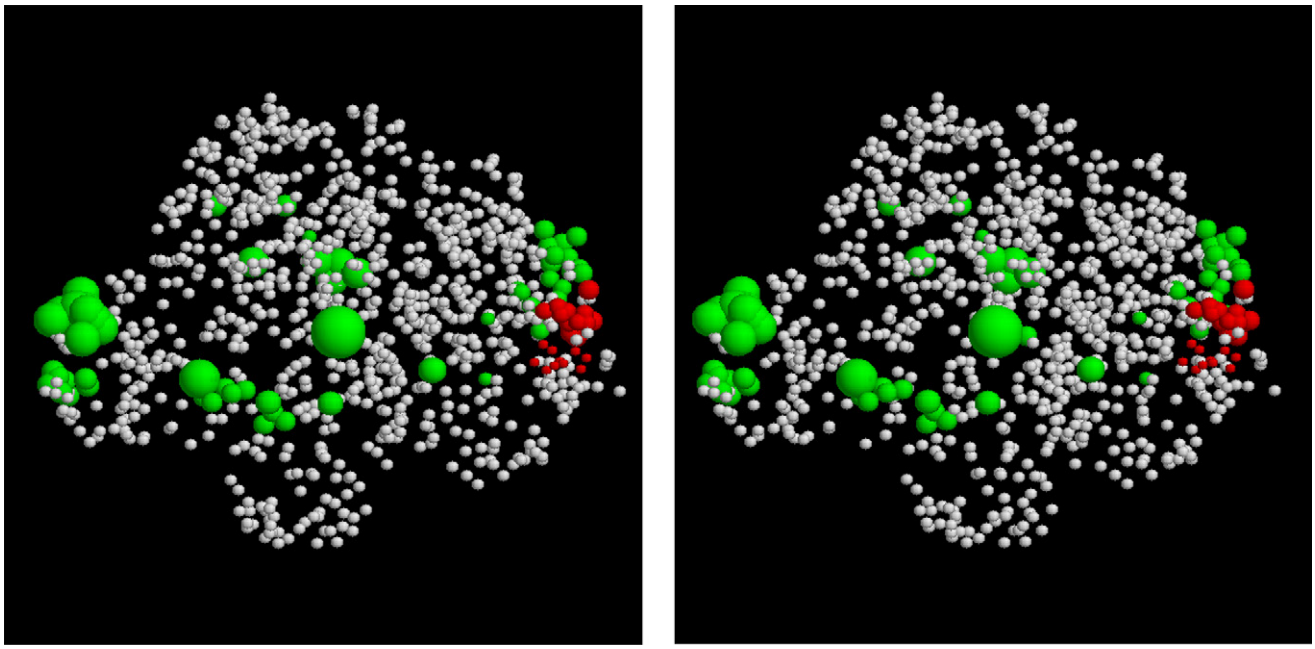


Figure 2. Protein Fold Space around 2trxA

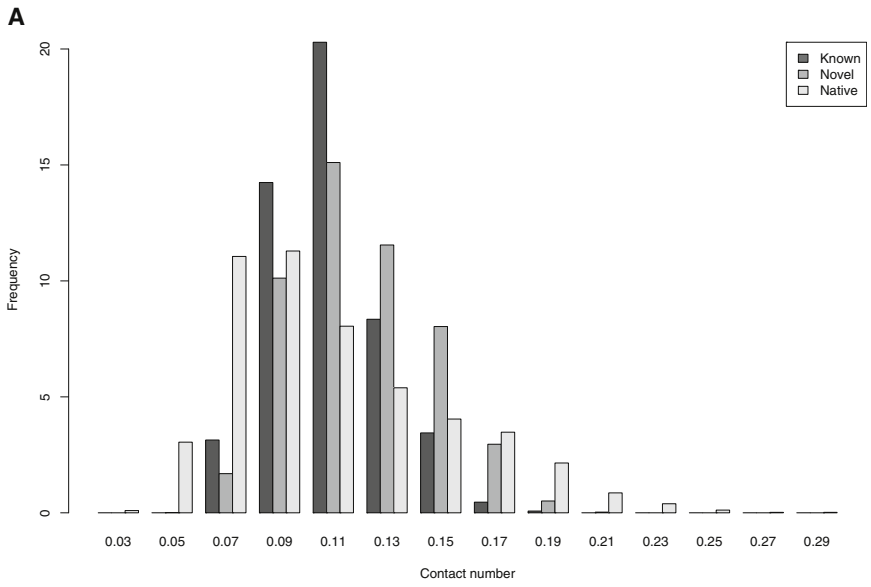
The distance between models was expressed as a metric and projected into 3D (see Experimental Procedures for details) for all pairs of folds generated with the thioredoxin starting probe (2trxA). Novel folds appear as small white spheres and known folds as colored spheres with a radius representing the number of proteins found to contain the fold. Red spheres mark those with close topology to the probe structure. The space is visualized as a stereo pair.

This unexpected number of novel folds was based on the correspondence of topological strings, and it is important to cross-check this result against more conventional methods of protein structure comparison. We did this by taking the α-carbon-based model of each model fold and scanning it against the nonredundant PDB using three standard protein structure comparison methods. These methods rely on the rmsd between the two structures as a measure of similarity, and this is known to be insensitive to topological rearrangements of strands in the β sheet (Taylor et al., 2008). (See Experimental Procedures for an example of this effect using the models shown in Figure 1). Therefore, each match of a novel model fold to a known protein with a low rmsd over most of the residues in the structures was examined by eye to decide whether the structures were topologically equivalent. We examined the best 200 of these structure matches obtained for each of the five proteins to ensure that no obvious similarities had been missed. In these 1000 comparisons, only 13 were found that could be considered a true topological match. We next considered the best match for each fold, with the models being ranked by their model quality score. After examining 200 of these comparisons, only two true topological matches were found. We considered this 1% error rate to be an acceptable level of accuracy for the comparisons based on the topology string comparisons.

The remaining list of just over 2000 novel folds was then analyzed for any features that might have made them distinct from either the models of known fold or from the native folds themselves. The distribution of numbers of secondary structure elements and their composition showed no great variation. This was not unexpected because these properties were derived from the predicted secondary structures through natural variation found among members or each sequence family. It was considered more likely that the novel folds might have been unobserved because they are more complex than those found in nature. This aspect was investigated using several measures including knot topology (Taylor, 2000), contact order (Plaxco et al., 1998), topological accessibility (Taylor, 2006), and some simpler values (described below) that could be extracted directly from the topology strings.

A measure of strand adjacency in each β sheet was calculated as the root mean square of the relative sheet positions between sequential β strands. These had the ratios 1.6: 1.8: 2.0 for known: novel: native,[2] showing some increase in the novel-fold models over the known-fold models but with the novel folds still less complex on average than native proteins by this measure. A similar simple measure was the number of parallel to antiparallel relationships between sequentially adjacent elements. A count over these pairs revealed a doubling in the percentage of parallel pairs in the novel-fold models over known-fold models but with the novel folds remaining in close agreement with the native structures (9.9: 20.2: 19.6 = known: novel: native). Because parallel packing

[2]The "known" group consists of models that have a fold that is found in known structures, "novel" is the group of models with no known corresponding fold and "native" are the folds observed in the non-redundant SCOP database of structures. To ensure that this group contained folds of comparable size, the native folds included only those domains that had a match to one of the three-layer αβα ideal Forms and the various measures were calculated only to the part of the structure covering this core. (See Experimental Procedures for full details.)



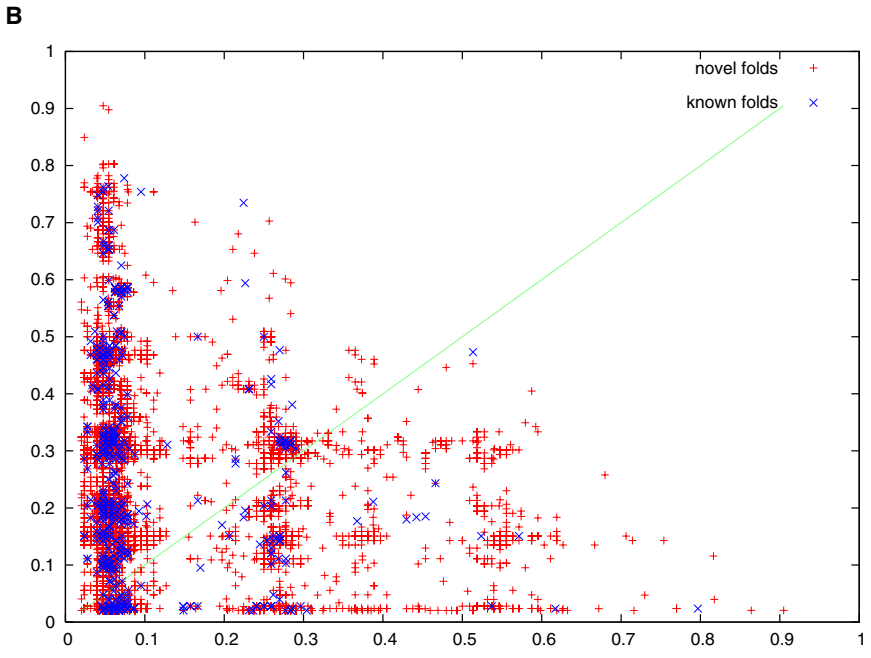


Figure 3. Fold-Complexity Measures

(A) The distribution of contact order over the known, novel, and native folds.

(B) The distribution of topological accessibility over the known and novel folds. Those with more local packing toward the amino-terminal end appear above the diagonal (green line) and simpler folds lie closer to the origin.

interactions are only forbidden in the core of our models, the novel folds appear to have explored the freedom allowed by this lack of constraint on edge connections—but not to the extent of generating un-protein-like models. When broken down by secondary structure type, only the β-β connections showed a marked deviation from native folds, but this was toward fewer parallel connections.

The chain contact order, which measures the sequential adjacency of packing, was calculated for all models and native structures and the distributions compared (Figure 3). This showed a slightly higher contact order for the novel-fold models relative to the known-fold models, but both had a smaller range than seen in native structures. The topological accessibility measure calculates the range of sequence over which the protein fold can be regenerated using only local contacts. Although related to the previous measures, this measure gives the additional perspective of whether the folds contain an amino-terminal "folding" bias, which is a strong feature in the βα folds (Taylor, 2006). The novel-fold models contained many folds that did not have an amino-terminal bias in their construction, but these derived simply from the greater total number of novel folds and the ratio of amino/carboxy bias was close to 2 for both the known and novel fold models (Figure 3B). That this unusual bias was preserved in the novel folds is a consequence of the multiple sequence alignment being used to construct the models. The majority of the five starting proteins have a buried amino-terminal β strand (typical for their class) and the conserved hydrophobicity associated with this region of sequence provides a bias for it to be buried in the models as well. We also tested each model fold for knots (Taylor, 2000) and found about 5% of them to be knotted. None of these were more complex than a simple trefoil, which is not unexpected for the size of protein chain considered (Taylor, 2007a). Although knots appear to be avoided by native proteins, folding studies suggest that this is not because of difficulty in folding (Mallam and Jackson, 2005), so we saw no reason to exclude them from our list of novel folds.

DISCUSSION

The specification of a protein fold is typically viewed as poorly defined, being sensitive to secondary structure elements that may be based on marginal or unconserved hydrogen bonds. Our use of a secondary structure lattice largely overcomes this problem. For models, the structure is built directly from the lattice and so their fold is predefined. For known structures, the ideal lattice must be matched to the structure, and although this step might appear susceptible to marginal effects, we do not

rely on a single best fit but retain all good fits to different lattice cores. This does not provide a unique definition of a protein fold, but in our study a match to any of the variations counts as a hit, giving a "fail-safe" position toward matching known folds. With these predefined fold definitions encoded as topology strings, there is no ambiguity in whether a pair of folds is the same because they will either have identical strings or strings that contain mismatches. This approach avoids the problems encountered using rmsd-based measures associated with length differences and statistical significance. Although we used this more familiar measure as an extensive cross-check on our topological measure, it was clear that it could not distinguish many distinct topological changes against the background "noise" associated with secondary structure shifts and loop variation. The important aspect of our method is that a systematic definition has been applied equally to models and native structures. We also believe that our definition, which maintains the integrity of the β sheet and the alpha layers on either side of the sheet, is not far from what is commonly accepted as a fold for this class of protein.

Our approach to protein fold definition is purely operational and does not imply any fundamental prescription of what a fold should be. If we had adopted a coarser granularity, then we would have had fewer PDB folds and also fewer model folds, or a finer level would have given more of each. Because our study is comparative, the particular level should not matter much but the level of secondary structure seems a natural one to adopt. Given fold discretization at this level, the only limitation on the distortion of a protein fold that we impose is swapping the position or orientation of secondary structures in a layer (e.g., β sheet). This is embodied in our topological definition and seems to be a reasonable constraint: if strands were allowed to swap positions in the sheet (or helices in a layer), one quickly approaches the situation where all folds are the same. In the other direction, if packing relationships (e.g., twist and shear) cannot vary much, then each protein quickly becomes unique. To select a level of granularity might seem to imply that fold space must be discrete at this level, rather than more toward a continuum as has been suggested (Petery and Honig, 2009; Sadreyev et al., 2009). In relation to this distinction, our folds would constitute nodes in a network between which steps could be made either with ease or difficulty. If all steps are easy, then fold space is continuous; if most are difficult, then fold space becomes discrete. The probability of a step is an evolutionary problem that has been considered previously (Johannissen and Taylor, 2004) but has not been addressed in the current work.

Given our systematic definition of a fold, we have shown that the number of folds in a region of fold space can be expanded easily by an order of magnitude over those with known folds—generating models that, by standard fold evaluation measures, exhibit no features that would justify their exclusion from the world of natural folds. The starting structures that we selected were all typical compact globular proteins with no unusual features and lie in a region of fold space that should have been well explored through the course of evolution. This large excess of potential novel folds was associated with each protein probe, and if this remains true in general, then the space of possible folds will always be larger than the space of realized folds, in much the same way that sequence space is always larger than fold space. This suggests that the natural folds constitute only a fraction of those that could exist. We can see no reason for this restriction other than the reuse of the same fold motifs through the process of gene duplication and copying. This finding has implications both for protein structure prediction and protein design. For prediction, the large number of neighboring nonnative folds explains why true ab initio structure prediction is so difficult compared with methods that construct models with cores and large fragments of known structures. For the latter approach, the PDB already contains sufficient components to reconstruct almost any fold (Friedberg and Godzik, 2005; Zhang and Skolnick, 2005a; Zhang et al., 2006) and it can be expected that such methods will perform increasingly well as the number of known folds increases. For protein design, our results suggest that the specification of template structures for the design of folds not previously seen in nature should not be a difficult task, given the rich underlying pool of novel folds from which they might be drawn.

In this study we have merely probed at a few points into the "dark matter" of protein fold space, but this has been sufficient to show that there is a plethora of unseen novel folds. For smaller proteins it has been shown previously that fold space will be more completely covered by known folds (Taylor, 2002), but for larger proteins the same analysis suggests that the number of unexplored topologies will expand greatly. As protein size increases, our estimate of a 10-fold ratio of known to novel folds may well be very conservative. If so, it would seem likely that there would not be room in the combined genomes of life on Earth to hold such a variety of proteins; however, the universe is very big and, like dark matter, the bulk might exist elsewhere.

EXPERIMENTAL PROCEDURES

We have extended an existing protein structure prediction method (Taylor et al., 2008) to generate a wide variety of folds based on variation around a known starting protein (referred to as the probe). Compared to randomly generating folds, our approach has an advantage because the models that we generate have realistic combinations and lengths of secondary structures and retain good hydrophobic packing in the core. Importantly, this prediction method uses only idealised constructs that do not derive directly from known protein structures or fragments of them. As a consequence, this provides a build-in test as we can check that the native fold of the probe protein is regenerated without being trivially 'copied' from its known structure or a homologue.

The comparison of both model and native folds is a critical aspect of the work and for this it is necessary to have a topology-based method as methods based on RMSD cannot be trusted to distinguish topological variation unless the structures being compared are quite similar. For this aspect we adapted a method previously used for protein classification (Taylor, 2002) which is based on a secondary structure lattice. With a protein structure mapped over such a lattice, its topology can be represented as a string and used to search for matches over collections of other topology strings derived from either models or PDB structures.

Most of the methods used in this work have been described previously but they will be summarised here for ease of reference. (For collected background material on the methods used in this work, see [Taylor and Aszódi, 2005]).

Model Fold Generation

Secondary structure predictions

As a starting point, the prediction method takes a secondary structure element (SSE) definition. These were generated from a multiple sequence alignment

using two artificial neural net based methods (PsiPred [Jones, 2000] and TUNE [Lin et al., 2002]). The methods were seeded with each sequence in the alignment in turn, resulting in a set of slightly differing predictions. SSE predictions were automatically 'tidied' to remove or reconcile fragmented predictions. However, if there was ambiguity whether a mixed prediction should be alpha or beta, then both variations were retained in separate predictions. Similarly, with small gaps between two SSEs of like type, broken and fused variations were kept.

From an alignment of between 10 to 20 sequences, this protocol typically resulted in around 50 distinct prediction variations. To extend this, ten different multiple alignments were generated for each starting protein drawn from a large pool of related sequences identified from a sequence databank search (using PsiBlast [Altschul et al., 1997]).

This approach generated secondary structure variations that remained consistent with the underlying nature of the sequence alignment, allowing this to be used as a basis on which the quality of the resulting models could be assessed.

Combinatorial fold generation

Each secondary structure sequence was used to generate compact fold topologies by a combinatorial search over a simple 2D-lattice. The lattice represented three layers of secondary structure with α-helices packed on either side of a β-sheet. Only connections between adjacent layers were allowed and these could not cross an existing connection on the same face (front or back). Parallel connections were only permitted between SSEs that lay on the edge of the lattice, otherwise in 3D a segment of chain would pass through the core of the model without any hydrogen bonds being formed (which is very unfavourable) [Taylor et al., 2008].

Alpha-carbon model construction

Each fold specified on the 2D-lattice was expanded into 3D as a bundle of twisted line-segments (sticks) representing the axis of each SSE. These in turn were elaborated as residue-level models with each amino acid (residue) being represented by its α-carbon position. SSEs were constructed with ideal geometry and their connecting loops by closed random walks.

The multiple sequence alignment was then aligned over this ideal α-carbon template, allowing conserved hydrophobic positions to co-locate in the core of the protein while improving the match of predicted and constructed secondary structure assignments. The model was finally adjusted to relieve steric clashes and refine local geometry. Our previous work used fragments of known structures to refine the models but even though this does not alter the fold of the starting model, we have omitted it from the current work to avoid any explicit link to known structures.

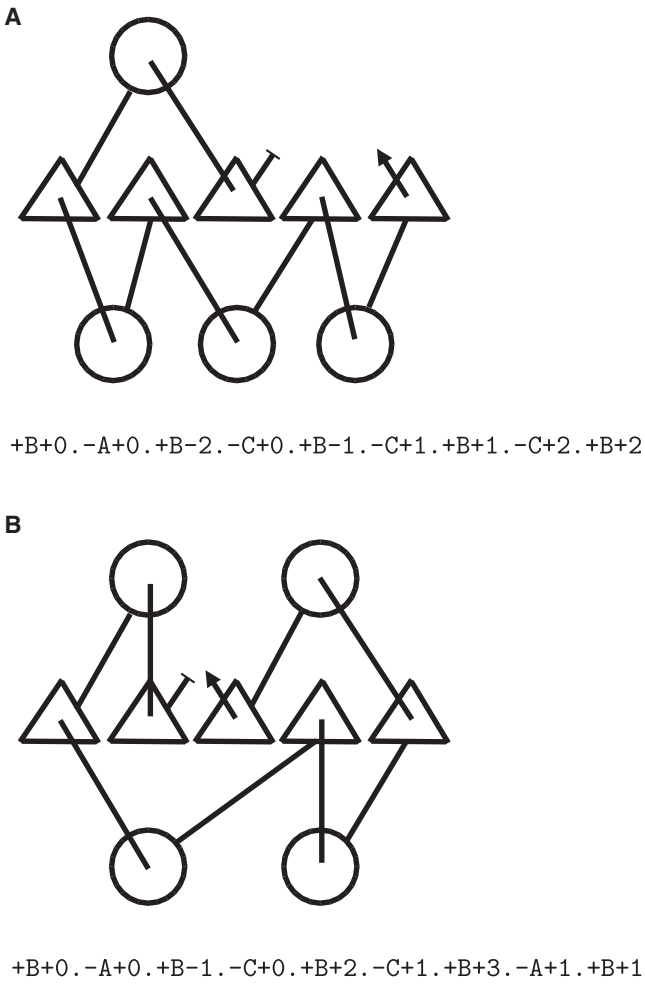


Figure 4. Example Topology Strings

Two small αβα layer proteins fitting form 1-5-3 (A) (one helix above and three below a five-stranded sheet) and form 2-5-2 (B) are shown as topology diagrams with their corresponding topology strings below. In the topology diagrams, helices are depicted as circles and β strands as triangles. In the topology strings, the three layers of secondary structure (αβα) are designated A (top), B, and C, respectively. Each secondary structure element (SSE) is given a label of three parts indicating orientation (+, −), layer, and position in the layer. The first SSE in each layer is, by definition, at position 0 with others numbered relative to this. In the topology diagram, negative numbers lie to the left and positive numbers to the right. Similarly, in the strings, a positive orientation corresponds to a SSE approaching ("out of the page") in the diagrams.

Model to Native Comparison

Protein fold definition

To characterize fold space, it is necessary to have a robust description of a protein fold that can be calculated automatically. Fortunately, this is already implicit in the model folds as they are all generated from a discrete path over an ideal lattice and have their fold specified by definition. However, the situation is not so simple for known (native) protein structures, against which the model folds will be compared. A problem is encountered as there exists some ambiguity over what constitutes a fold and at which point pairs of structures can be considered to have distinct folds [Kolodny et al., 2006; Taylor, 2007b]. For consistency, we adopted an approach based on the comparison of native structures against the same idealised secondary structure lattices over which the model folds are generated. Previously a similar approach was used to classify known structures into a system that had some similarity to the Periodic Table of elements [Taylor, 2002]. This earlier study concentrated on finding the best fit of an intact protein to the lattice models (called Forms) whereas the current task requires the more general problem of whether a model protein corresponds to any compact part of a native protein.

Focusing on globular protein domains that contain a single β-sheet and do not form a circular β-barrel, we matched every structure in the non-redundant SCOP-Astral domain collection against every compact fragment of the ideal lattice from three strands up to thirteen strands in the β-sheet. In addition, all variants were considered with no α-helices packed on the sheet up to the maximum possible packed on both sides. For each structure, we did not select a single best match but retained all reasonable matches. This means that if there is a known structure in which it is ambiguous whether, say, an α-helix on the edge of the domain should be considered as part of the fold, then there will exist two entries in the list of fold definitions: one in which the helix is included and one where it is omitted. When a model fold is matched against these definitions, it will be free to match either of these ideal Forms. This provides a robust way to test if any model fold specified from the lattice has a correspondence in the real world.

Topology string encoding and matching

The folds for both the native and model structures were encoded as strings that specify the coordinates of the chain through the lattice. Each match to a Form allows the fold to be specified by its path over the underlying lattice. This can be done in a simple coordinate system which uses the letters "A", "B" and "C" for the three layers, and a number for position in the layer with

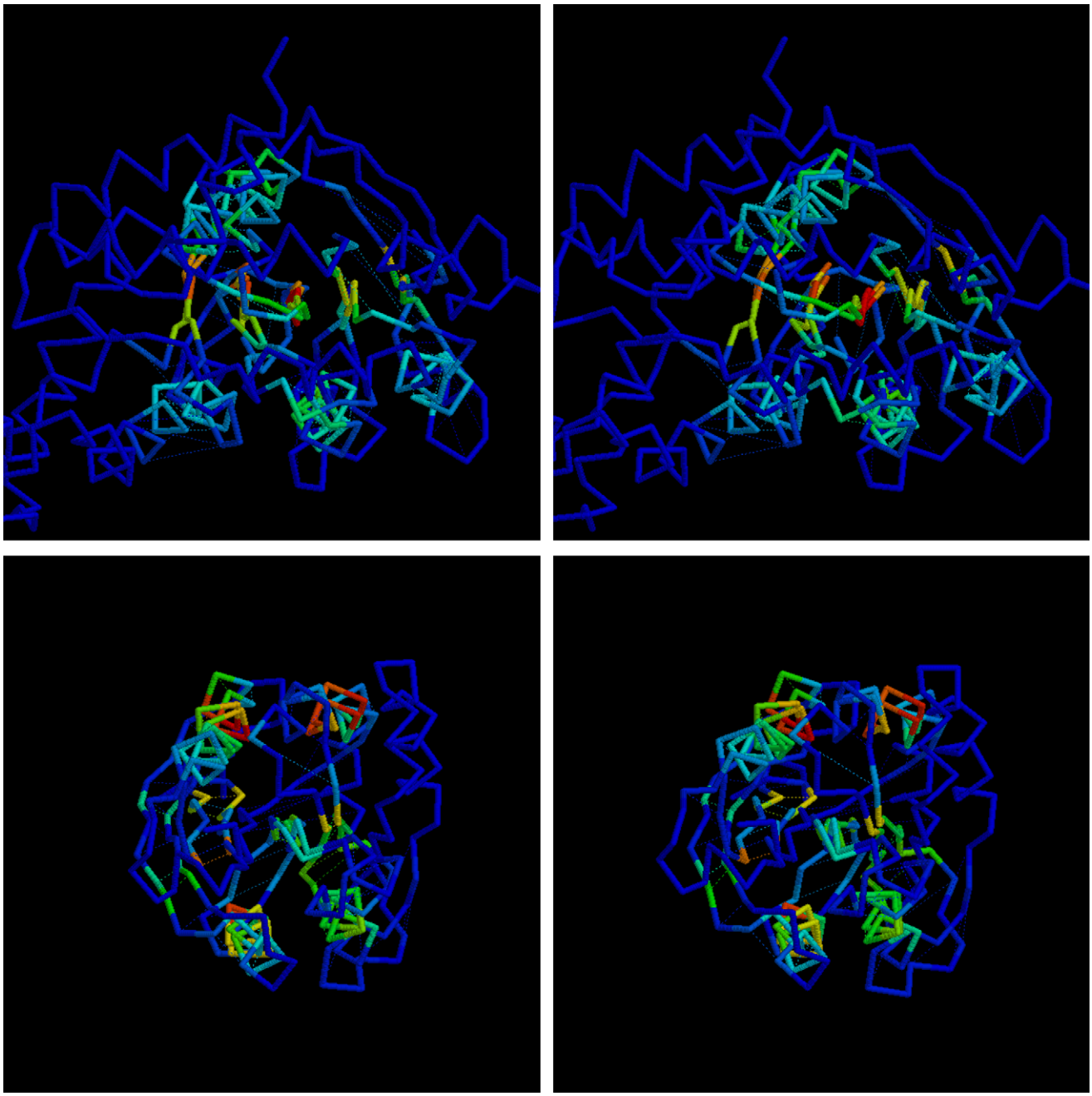


Figure 5. Example Rmsd-Based Comparisons

A model with a known fold (top frames) and one with a novel fold (lower frames) are shown superimposed (as a stereo pair) on the best match obtained to a native protein structure in the PDB. The structural alignment and superposition was calculated by the program SAP and the residues are colored according to their degree of structural correspondence (red = best to blue = none). The match to 1gnvB (top) has a full topological correspondence over the model but with unmatched additional helices in the PDB structure. The match to 1uuyA (lower) has a topological mismatch of strands in the core β sheet, despite having an equally good rmsd fit. (Compare with Figures 1 and 4, which depict the same folds, and see text for discussion).

the remaining dimension requiring only two values, "+" or "-", to designate front and back. The first SSE to enter a layer is assigned position 0 and the first strand in the sheet takes the positive orientation, giving "+B+0" in the string. The first α-helix then sets the top/bottom orientation by assigning its layer as "A". The resulting strings (referred to as "topology strings") are quite easy to read and visualise the fold. Two examples are given in Figure 4 corresponding to the example folds illustrated in Figure 1. The first (A) is a common fold and has 29 matches in the non-redundant domain database whereas the second (B) has none and is classed as a novel fold.

To allow for substructure matching, each string was progressively truncated from the beginning with the SSEs being renumbered accordingly. Any truncated string that did not encode a compact β-sheet with more than two strands was removed. For example in Figure 4, removal of the first strand leaves only the last two whereas in Figure , a compact 3-stranded structure remains with

topology string: +B+0.-C+0.+B+1.-A+0.+B-1. To maintain the orientation of the fold after truncation, with the first strand as "+B+0" and the first helix in the "A" layer, typological manipulations (*e.g.* flipping AC and signs), can be used to 'rotate' the fold. The previous string thus becomes: +B+0.-A+0.+B-1.-C+0.+B+1[1]. All reoriented substrings were stored to allow standard string matching methods (e.g., the Linux utility grep) to check if a model fold exists among the known structures.

Structure-based matching

Comparison of the α-carbon models to native structures was made using three different structure comparison methods including DALI [Holm and Sander, 1993], TMalign [Zhang and Skolnick, 2005b] and SAP [Taylor, 1999]. The first two concentrate on finding well-matched cores and subsets of positions below a cutoff distance, respectively. This was found to result often in a good match that did not cover the full structure (data not shown). As we are interested in overall topology, substantial segments of the structure cannot be omitted. To avoid partial matches, we concentrated on the SAP method, which is not strictly distance based, in combination with a filter to discard partial matches. Rather than use a sharp length cutoff, a sigmoidal switch function was used to damp the scores of partial matches: $1/(1+\exp((0.7*N-x+10)/5))$, where N is the length of the protein and x is the number of matched positions. Matches under 90% of N are downweighted by 0.8 and those under 80% lose half their match score. Even with this filter, considerable visual inspection was still required.

The necessity for visual inspection when relying on RMSD-based measures is illustrated retaining the two example folds used above in Figure and Figure . Good matches with PDB structures were obtained for the model with a known fold (Figures 1 and 4) and its superposition on the protein with PDB code 1uuyA is shown in Figure 5 (top). This match corresponds to the core fold of 1uuyA and the additional helices (seen in dark blue to the front in the stereo image) are unmatched. The core has a RMSD fit of 4.8Å over 107 equivalent α-carbon atoms. However, the model with a novel fold has an almost equally good fit of 5.1Å RMSD over 106 α-carbon atom pairs to the protein with PDB code 1gvnB (Figure 5, lower) — despite having a mismatched topology. Such a small difference in RMSD fit cannot be distinguished against background noise and, without a topology-based measure, it would be necessary to visually examine every superposition. Even with inspection using interactive graphics, topological differences are not always apparent. In this example it can only be seen with difficulty in the superposition (Figure 5) that the good RMSD fit to 1gvnB is obtained by equivalent locations being preserved for the α-helices but with the central β-strand being skipped and the remaining four strands in the model fitting in the gaps between the strands in the PDB structure. Smaller topological changes, such as the swapping of two adjacent strand positions in a β-sheet can easily be missed in a structural superposition but, of course, cause a clear mismatch in a topology string comparison.

Fold-space projection

The relationships between sets of protein folds would ideally be viewed as a function of their topology alone but as there is no simple metric to relate topologies, an RMSD measure was used but modified by taking the square-root of the RMSD when the topology strings were the same.

These values were visualised by using a gradual multi-dimensional projection method with triangle-inequality violation correction and multidimensional Euclidean distance refinement in high dimensions until an three-dimensional solution could be obtained [Aszódi et al., 1995]. Direct projection in 3D using a simple principle component analysis algorithm did not result in a meaningful representation.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures and can be found with this article online at http://www.cell.com/structure/supplemental/S0969-2126(09)00295-0.

[1]Comparing this to the full string it can be seen immediately that this fold contains an internal structural repeat represented by the common substring: +B+0.−A+0.+B−1.−C+0 consisting of the first two β-strands and their flanking helices.

ACKNOWLEDGMENTS

Most of the calculations for this work were made on the NOTUR facility at the University of Tromso using the 5000 CPU STALLO computer cluster. The work was supported by the Medical Research Council (UK) and the National Programme for Research in Functional Genomics in Norway (FUGE) in the Research Council of Norway. J.T.M. was supported by DARPA.

Received: March 2, 2009
Revised: July 13, 2009
Accepted: July 14, 2009
Published: September 8, 2009

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Aszódi, A., Gradwell, M.J., and Taylor, W.R. (1995). Global fold determination from a small number of distance restraints. J. Mol. Biol. *251*, 308–326.

Friedberg, I., and Godzik, A. (2005). Connecting the protein structure universe by using sparse recurring fragments. Structure *13*, 1203–1211.

Grant, A., Lee, D., and Orengo, C. (2004). Progress towards mapping the universe of protein folds. Genome Biol. *5*, 107.

Holm, L., and Sander, C. (1993). Protein-structure comparison by alignment of distance matrices. J. Mol. Biol. *233*, 123–138.

Johannissen, L.O., and Taylor, W.R. (2004). Protein fold comparison by the alignment of topological strings. Prot. Eng. *16*, 949–955.

Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404–405.

Kolodny, R., Petery, D., and Honig, B. (2006). Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Curr. Opin. Struct. Biol. *16*, 393–398.

Lin, K., May, A.C., and Taylor, W.R. (2002). Threading using neural networks (TUNE): the measure of protein sequence-structure compatibility. Bioinformatics *18*, 1350–1357.

Mallam, A.L., and Jackson, S.E. (2005). Folding studies on a knotted protein. J. Mol. Biol. *346*, 1409–1421.

Petery, D., and Honig, B. (2009). Is protein classification necessary? towards alternative approaches to function prediction. Curr. Opin. Struct. Biol. *19*, 1–6.

Plaxco, K.W., Simons, K.T., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. *277*, 985–994.

Sadreyev, R.I., Kim, B.-H., and Grishin, N.V. (2009). Discrete-continuous duality of protein structure space. Curr. Opin. Struct. Biol. *19*, 321–328.

Taylor, W.R. (1999). Protein structure alignment using iterated double dynamic programming. Protein Sci. *8*, 654–665.

Taylor, W.R. (2000). A deeply knotted protein and how it might fold. Nature *406*, 916–919.

Taylor, W.R. (2002). A periodic table for protein structure. Nature *416*, 657–660.

Taylor, W.R. (2006). Topological accessibility shows a distinct asymmetry in the folds of βα proteins. FEBS Lett. *580*, 5263–5267.

Taylor, W.R. (2007a). Protein knots and fold complexity: some new twists. Comput. Biol. Chem. *31*, 151–162.

Taylor, W.R. (2007b). Evolutionary transitions in protein fold space. Curr. Opin. Struct. Biol. *17*, 354–361.

Taylor, W.R., and Aszódi, A. (2005). Protein Geometry, Classification, Topology and Symmetry: A Computational Analysis of Structure (CRC Press).

Taylor, W.R., Bartlett, G.J., Chelliah, V., Klose, D., Lin, K., Sheldon, T., and Jonassen, I. (2008). Prediction of protein structure from ideal forms. Proteins *70*, 1610–1619.

Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., and Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. Proc. Natl. Acad. Sci. USA *103*, 2605–2610.

Zhang, Y., and Skolnick, J. (2005a). The protein structure prediction problem could be solved using the current PDB library. Proc. Natl. Acad. Sci. USA *102*, 1029–1034.

Zhang, Y., and Skolnick, J. (2005b). A protein structure alignment algorithm based on TM-score. Nucleic Acids Res. *33*, 2302–2309.

